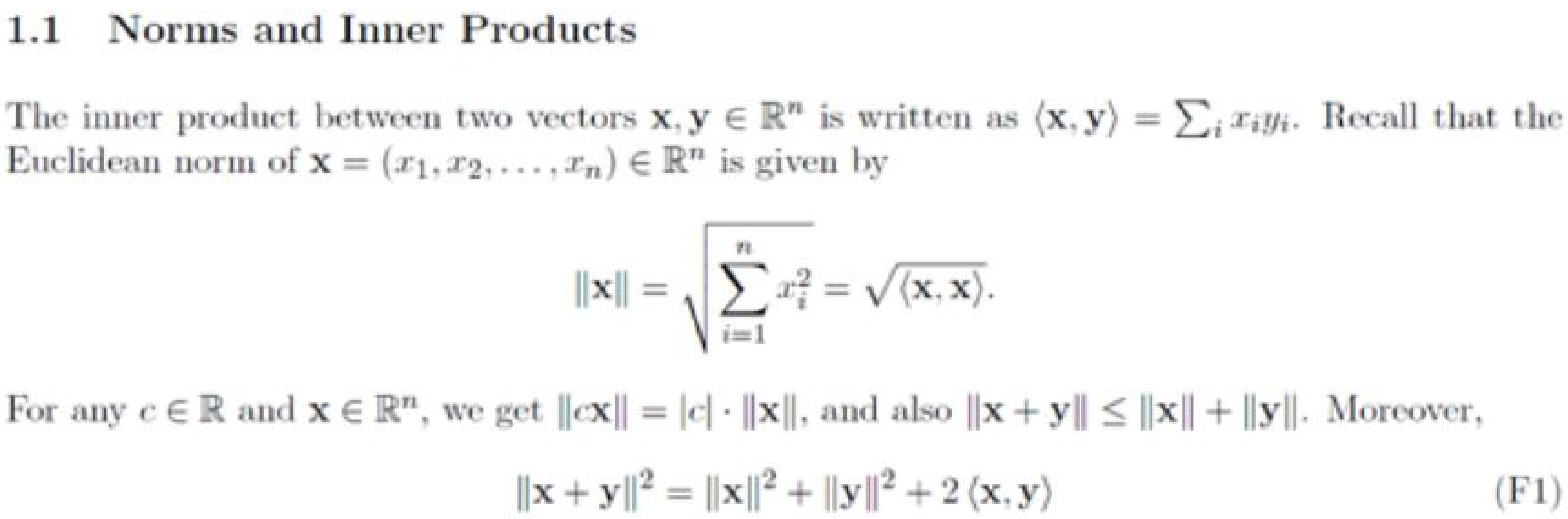
Gradient Descent



1.1 Norms and Inner Products

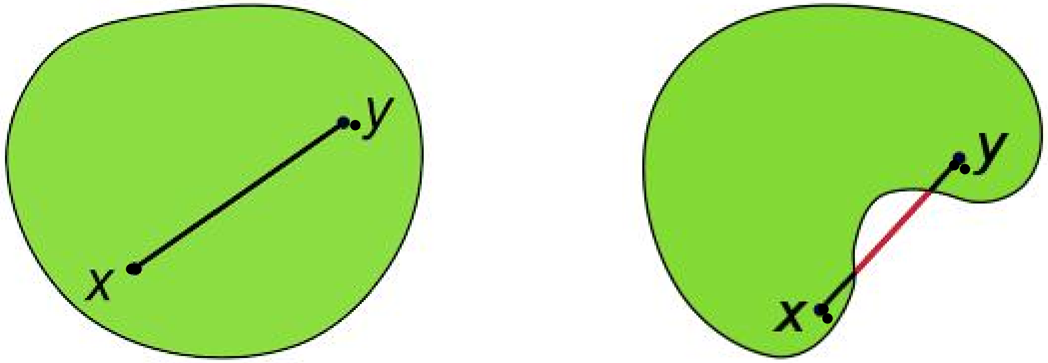
The inner product between two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ is written as $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_i x_i y_i$. Recall that the Euclidean norm of $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ is given by

$$\|\mathbf{x}\| = \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}.$$

For any $c \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^n$, we get $\|c\mathbf{x}\| = |c| \cdot \|\mathbf{x}\|$, and also $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$. Moreover,

$$\|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + 2\langle \mathbf{x}, \mathbf{y} \rangle \tag{F1}$$

Convexity



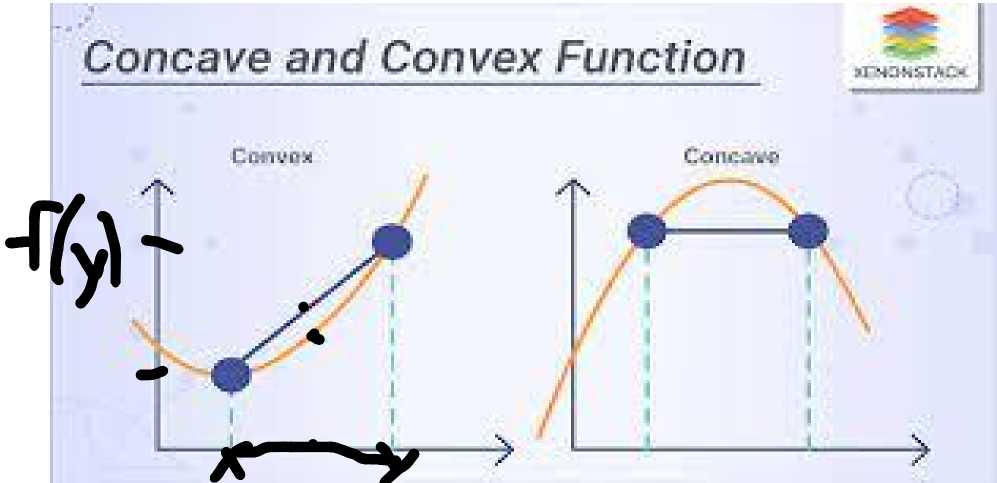
Definition 1 *A set $K \subseteq \mathbb{R}^n$ is said to be convex if*

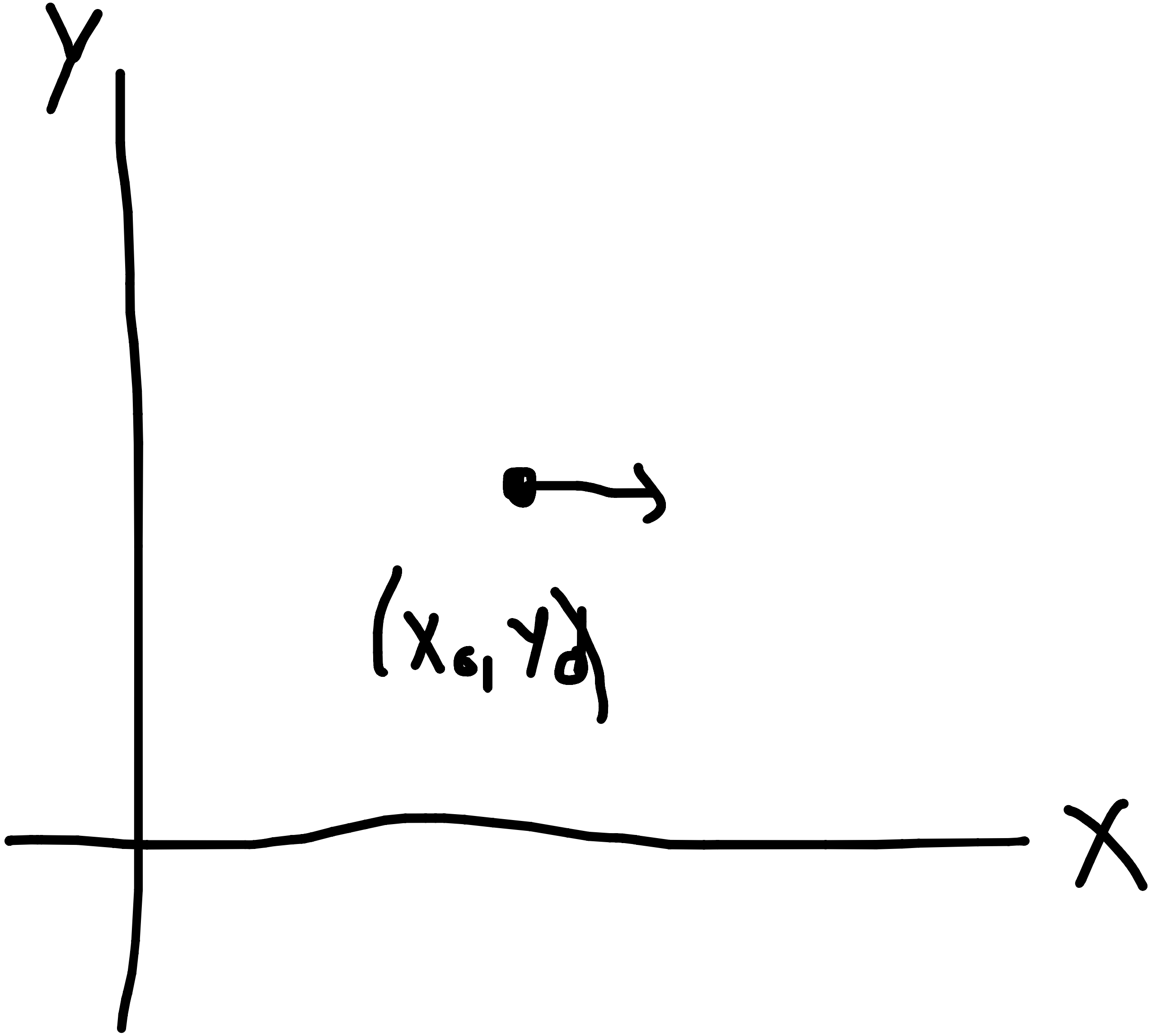
$$(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \in K \qquad \forall \mathbf{x}, \mathbf{y} \in K, \forall \lambda \in [0,1]$$

Definition 2 *For a convex set $K \subseteq \mathbb{R}^n$, a function $f : K \to \mathbb{R}$ is said to be convex over K iff*

$$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \forall \lambda \in [0,1]$$

Whenever K is not specified, assume $K = \mathbb{R}^n$.





$$\left.\frac{\partial f}{\partial x}\right|_{(x_0,y_0)}$$

$$\left.\frac{\partial f}{\partial y}\right|_{(x_0}$$

$$f(x_0+\varepsilon, y_0+\delta) = f(x_0, y_0) + \varepsilon \left.\frac{\partial f}{\partial x}\right|_{(x_0,y_0)} + \delta \frac{\partial f}{\partial y}$$

$\underline{f(x+v) \geq f(x) + \langle v, \nabla f(x) \rangle} \leftarrow v = y - x$

Gradients

In the context of this lecture, we will always assume that the function f is differentiable. The analog of the derivative in the multivariate case is the *gradient* ∇f, which is itself a function from $K \to \mathbb{R}^n$ defined as follows:

$$\underline{\nabla f(\mathbf{x})} = \left(\frac{\partial f}{\partial x_1}(\mathbf{x}), \ldots, \frac{\partial f}{\partial x_n}(\mathbf{x}) \right).$$

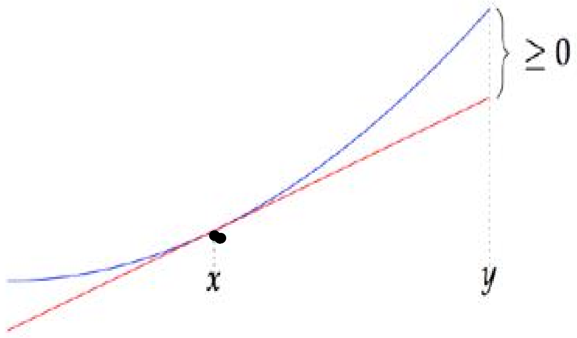
$\frac{\partial f}{\partial x \partial n}$

We assume that the gradient is well-defined at all points in K.[2] Visually, if you draw the "level sets" of points where the function f takes on the same value as at $\mathbf{x}$, then the gradient $\nabla f(x)$ gives you the tangent plane to this level set at point $\mathbf{x}$.

Fact 3 *A differentiable function $f : K \to \mathbb{R}$ is convex iff $\forall \mathbf{x}, \mathbf{y} \in K$,*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle .$$

$K = \mathbb{R}^n$

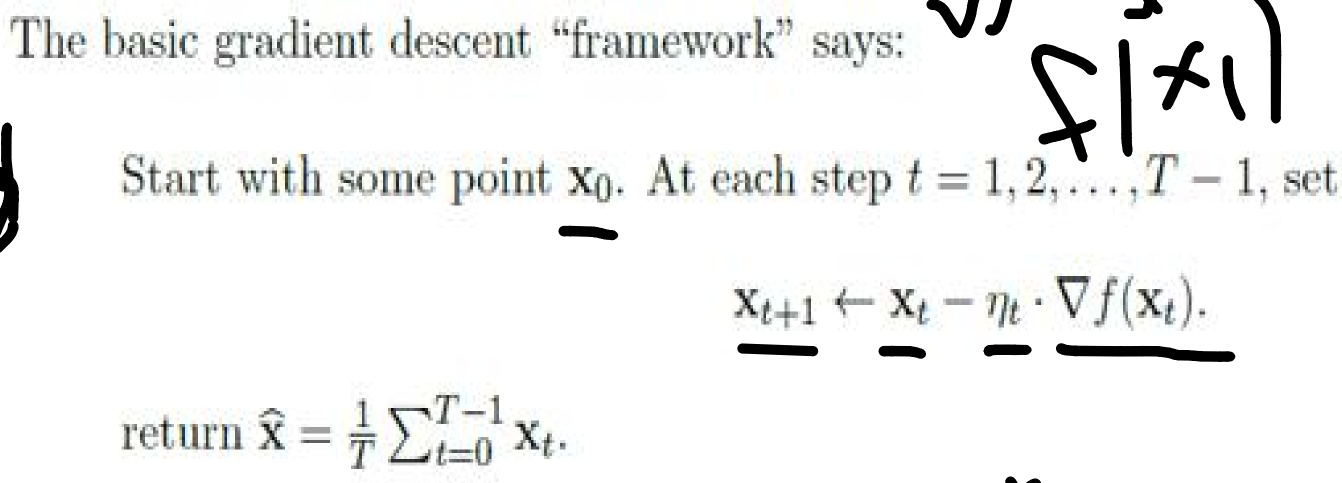


Minimizing a Function

- To minimize a function, set its gradient to equal 0
  - Finds global minimum if f is a convex function
  - For non-convex function, still often finds a good solution, i.e., a local minimum
- Gradient is a very complicated expression, can’t solve by setting to 0
- Instead, update iteratively. This is called gradient descent

Gradient Descent

$f(x_0)$ $\}$ ✓ $f(x_1)$

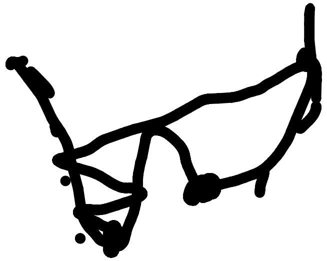


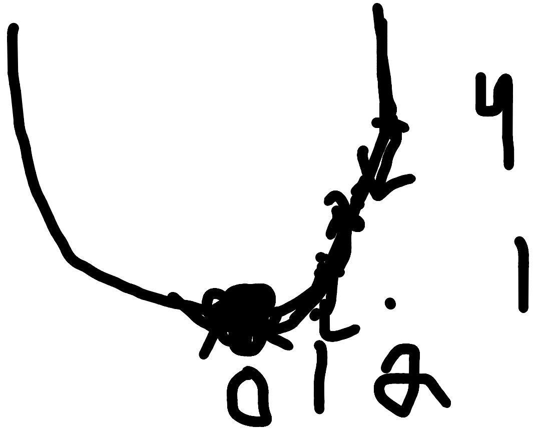
The basic gradient descent "framework" says:

Start with some point x_0. At each step $t = 1, 2, \ldots, T-1$, set

$$\mathrm{x}_{t+1} \leftarrow \mathrm{x}_t - \eta_t \cdot \nabla f(\mathrm{x}_t). \quad (1)$$

return $\hat{\mathrm{x}} = \frac{1}{T}\sum_{t=0}^{T-1} \mathrm{x}_t$.





$x_0 = 2$

$f(x) = x^2$

x_T

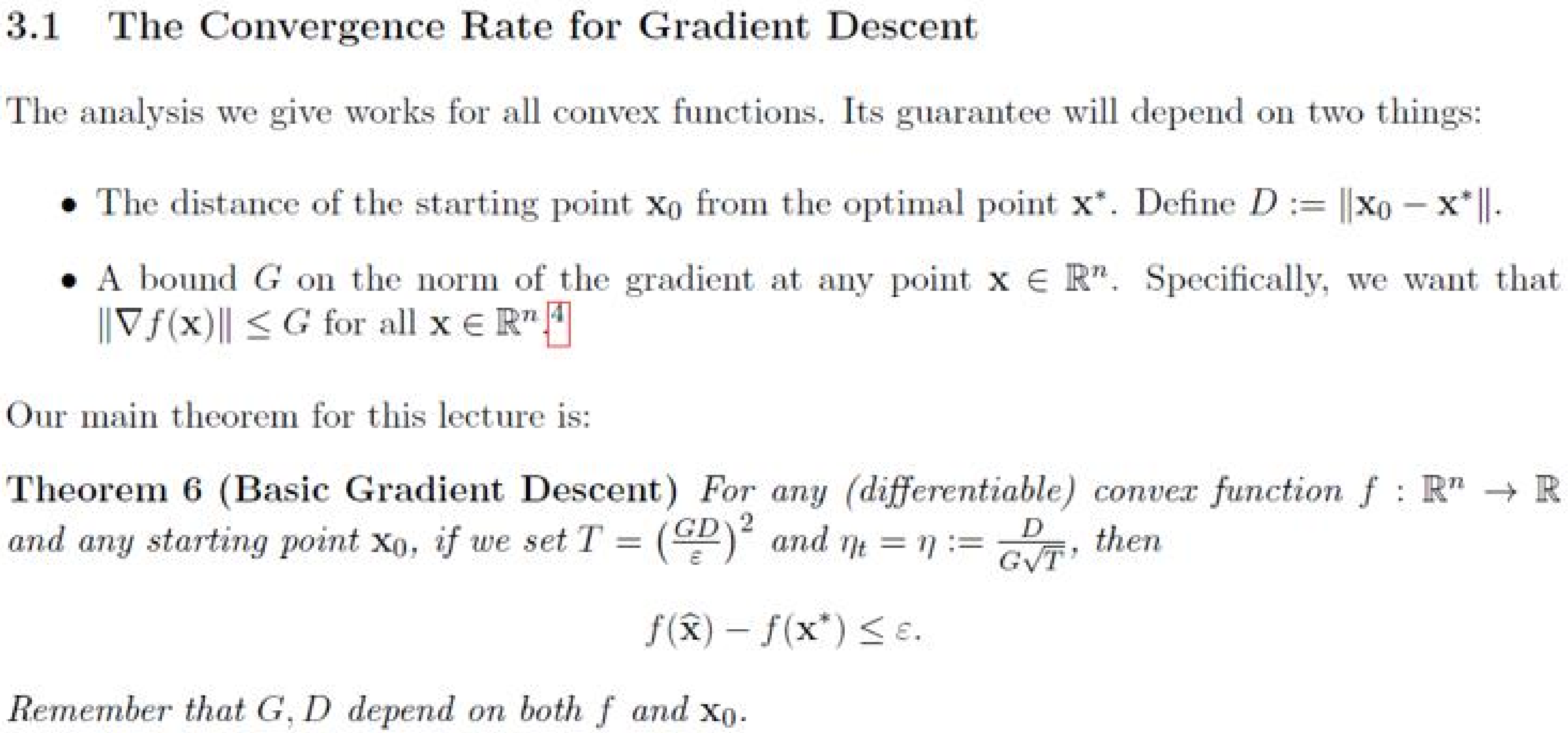
$x_1 = 2 - \frac{1}{4} \cdot 4$

$\frac{df}{dx} = 2x$

$x_1 = 1$

- η_t is a "learning rate" $\frac{1}{4}$
- We "move a little" in a direction (the negative gradient) that reduces loss function
  - Think of rolling ball down a hill
- Can just output x_t in practice, though often easier to prove statements about $\hat{\mathrm{x}}$

Gradient Descent Convergence



3.1 The Convergence Rate for Gradient Descent

The analysis we give works for all convex functions. Its guarantee will depend on two things:

- The distance of the starting point $\mathbf{x}_0$ from the optimal point $\mathbf{x}^*$. Define $D := \|\mathbf{x}_0 - \mathbf{x}^*\|$.
- A bound G on the norm of the gradient at any point $\mathbf{x} \in \mathbb{R}^n$. Specifically, we want that $\|\nabla f(\mathbf{x})\| \leq G$ for all $\mathbf{x} \in \mathbb{R}^n$.[4]

Our main theorem for this lecture is:

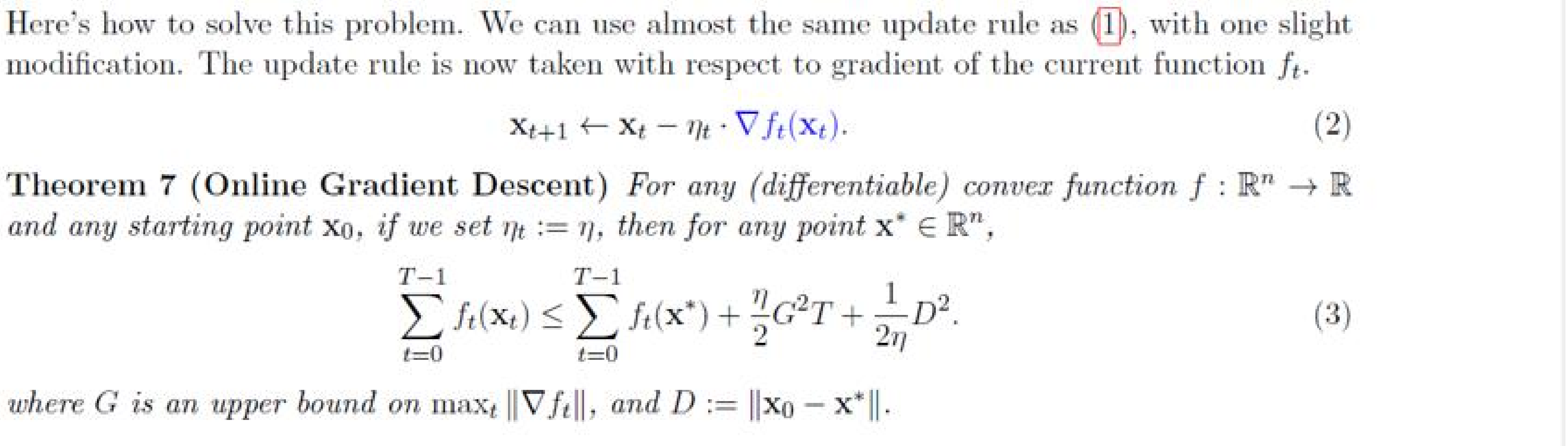
Theorem 6 (Basic Gradient Descent) *For any (differentiable) convex function $f : \mathbb{R}^n \to \mathbb{R}$ and any starting point $\mathbf{x}_0$, if we set $T = \left(\frac{GD}{\varepsilon}\right)^2$ and $\eta_t = \eta := \frac{D}{G\sqrt{T}}$, then*

$$f(\widehat{\mathbf{x}}) - f(\mathbf{x}^*) \leq \varepsilon.$$

Remember that G, D depend on both f and $\mathbf{x}_0$.

A Stronger Statement – Online Gradient Descent

- Suppose we even allow the function f_t to change at each time step



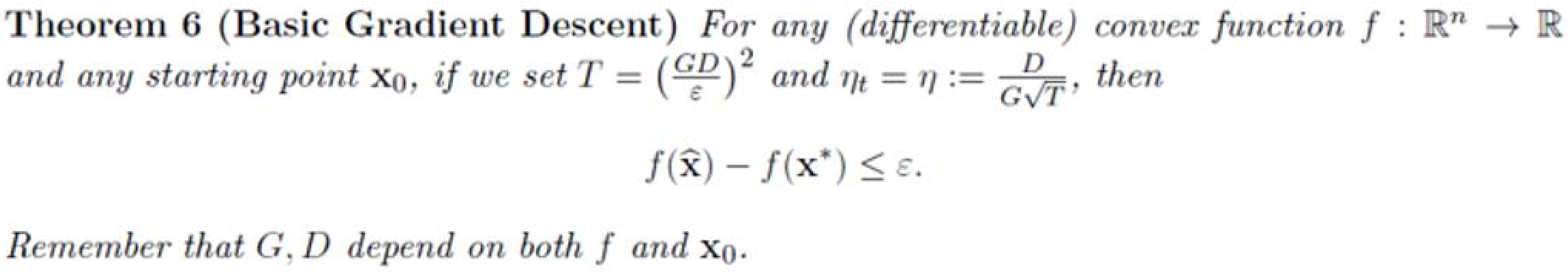
Here's how to solve this problem. We can use almost the same update rule as (1), with one slight modification. The update rule is now taken with respect to gradient of the current function f_t.

$$\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \eta_t \cdot \nabla f_t(\mathbf{x}_t). \tag{2}$$

Theorem 7 (Online Gradient Descent) *For any (differentiable) convex function $f : \mathbb{R}^n \to \mathbb{R}$ and any starting point $\mathbf{x}_0$, if we set $\eta_t := \eta$, then for any point $\mathbf{x}^* \in \mathbb{R}^n$,*

$$\sum_{t=0}^{T-1} f_t(\mathbf{x}_t) \leq \sum_{t=0}^{T-1} f_t(\mathbf{x}^*) + \frac{\eta}{2} G^2 T + \frac{1}{2\eta} D^2. \tag{3}$$

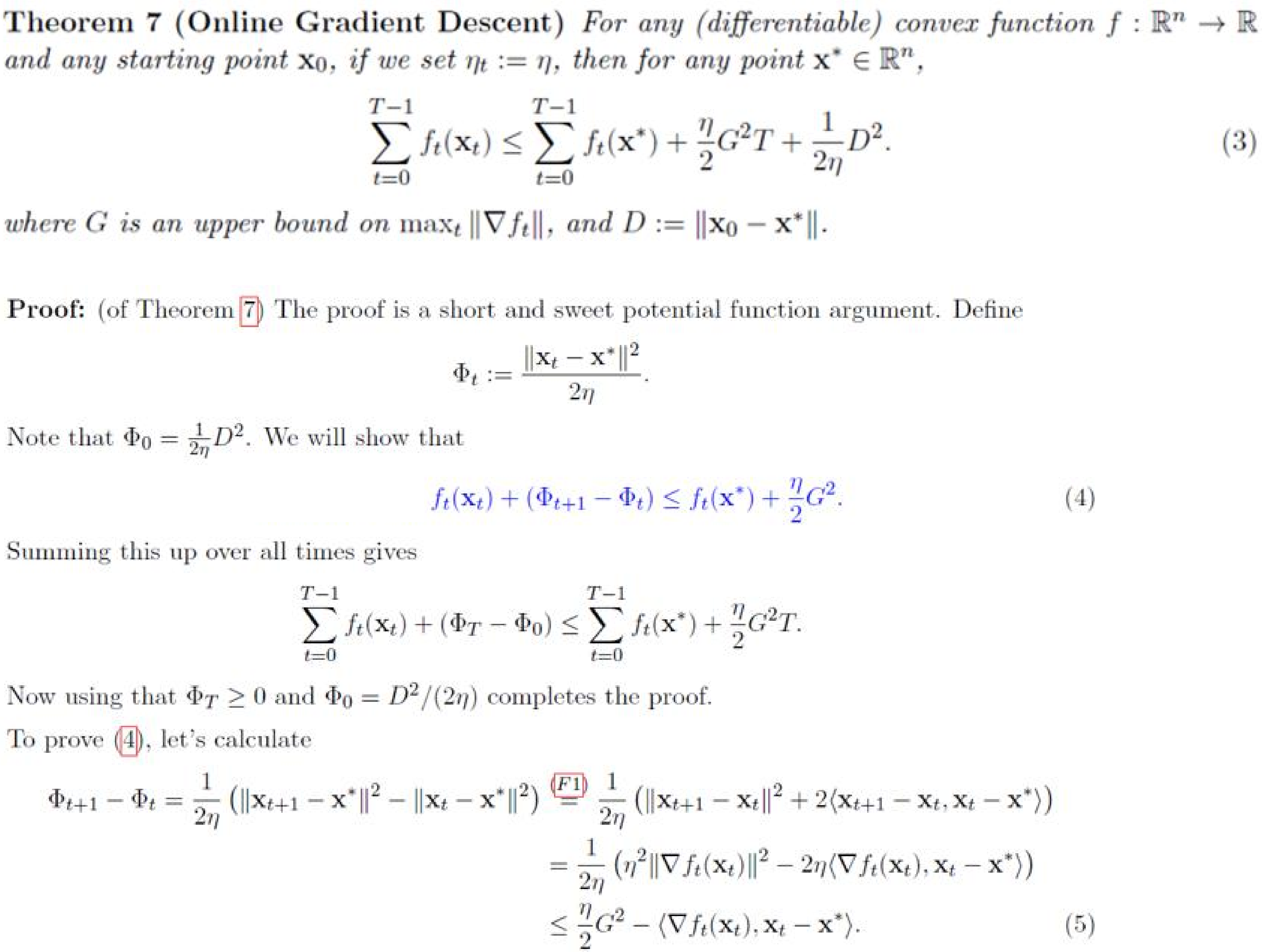
where G is an upper bound on $\max_t \|\nabla f_t\|$, and $D := \|\mathbf{x}_0 - \mathbf{x}^*\|$.



Theorem 6 (Basic Gradient Descent) *For any (differentiable) convex function $f : \mathbb{R}^n \to \mathbb{R}$ and any starting point $\mathbf{x}_0$, if we set $T = \left(\frac{GD}{\varepsilon}\right)^2$ and $\eta_t = \eta := \frac{D}{G\sqrt{T}}$, then*

$$f(\widehat{\mathbf{x}}) - f(\mathbf{x}^*) \leq \varepsilon.$$

Remember that G, D depend on both f and $\mathbf{x}_0$.



Theorem 7 (Online Gradient Descent) *For any (differentiable) convex function $f : \mathbb{R}^n \to \mathbb{R}$ and any starting point $\mathbf{x}_0$, if we set $\eta_t := \eta$, then for any point $\mathbf{x}^* \in \mathbb{R}^n$,*

$$\sum_{t=0}^{T-1} f_t(\mathbf{x}_t) \le \sum_{t=0}^{T-1} f_t(\mathbf{x}^*) + \frac{\eta}{2}G^2T + \frac{1}{2\eta}D^2. \qquad (3)$$

where G is an upper bound on $\max_t \|\nabla f_t\|$, and $D := \|\mathbf{x}_0 - \mathbf{x}^*\|$.

Proof: (of Theorem 7) The proof is a short and sweet potential function argument. Define

$$\Phi_t := \frac{\|\mathbf{x}_t - \mathbf{x}^*\|^2}{2\eta}.$$

Note that $\Phi_0 = \frac{1}{2\eta}D^2$. We will show that

$$f_t(\mathbf{x}_t) + (\Phi_{t+1} - \Phi_t) \le f_t(\mathbf{x}^*) + \frac{\eta}{2}G^2. \qquad (4)$$

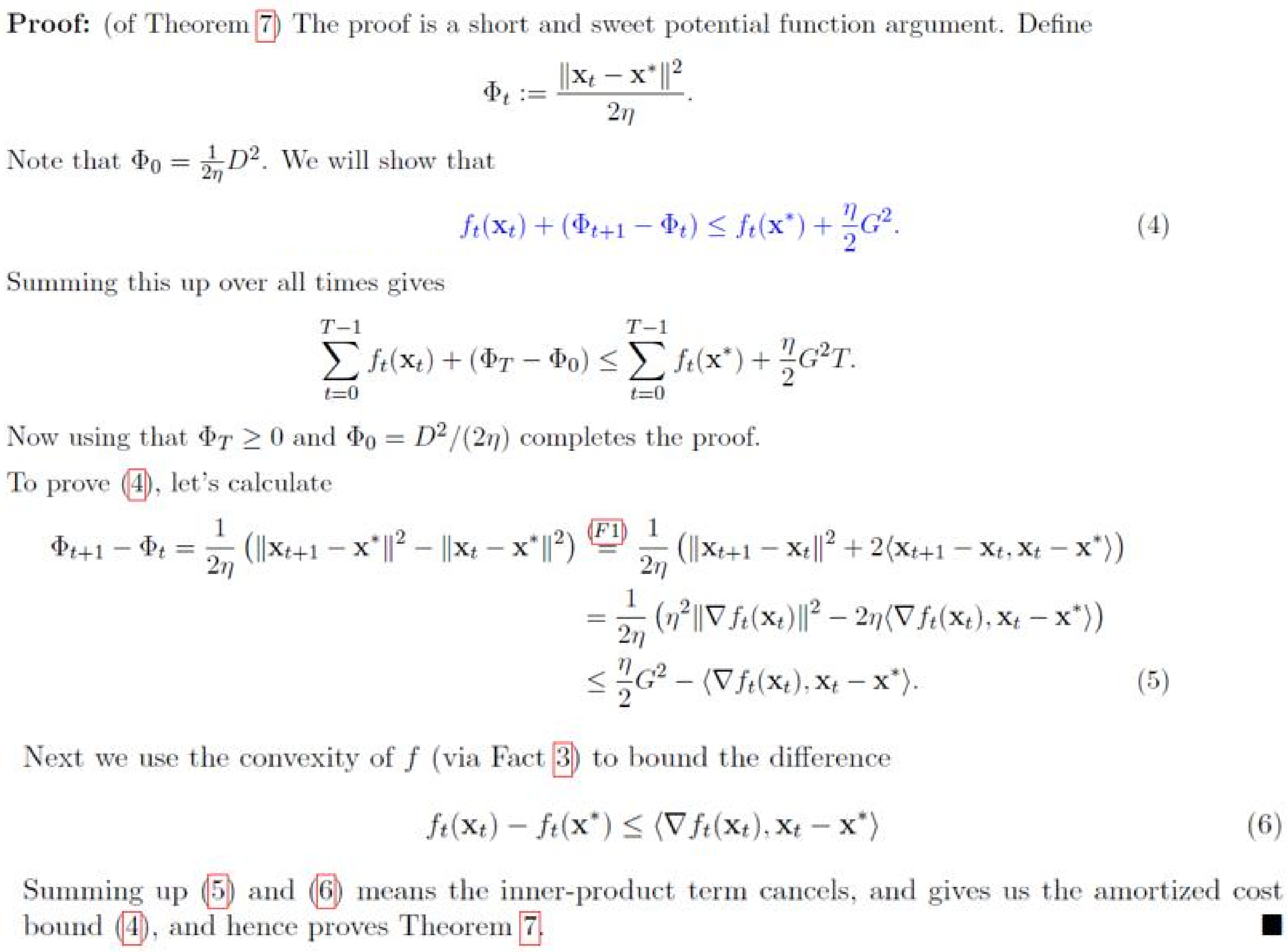
Summing this up over all times gives

$$\sum_{t=0}^{T-1} f_t(\mathbf{x}_t) + (\Phi_T - \Phi_0) \le \sum_{t=0}^{T-1} f_t(\mathbf{x}^*) + \frac{\eta}{2}G^2T.$$

Now using that $\Phi_T \ge 0$ and $\Phi_0 = D^2/(2\eta)$ completes the proof.

To prove (4), let's calculate

$$\begin{aligned}\Phi_{t+1} - \Phi_t = \frac{1}{2\eta}\left(\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 - \|\mathbf{x}_t - \mathbf{x}^*\|^2\right) &\overset{(F1)}{=} \frac{1}{2\eta}\left(\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + 2\langle \mathbf{x}_{t+1} - \mathbf{x}_t, \mathbf{x}_t - \mathbf{x}^*\rangle\right) \\ &= \frac{1}{2\eta}\left(\eta^2\|\nabla f_t(\mathbf{x}_t)\|^2 - 2\eta\langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^*\rangle\right) \\ &\le \frac{\eta}{2}G^2 - \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^*\rangle. \qquad (5)\end{aligned}$$



Proof: (of Theorem 7) The proof is a short and sweet potential function argument. Define

$$\Phi_t := \frac{\|\mathbf{x}_t - \mathbf{x}^*\|^2}{2\eta}.$$

Note that $\Phi_0 = \frac{1}{2\eta}D^2$. We will show that

$$f_t(\mathbf{x}_t) + (\Phi_{t+1} - \Phi_t) \le f_t(\mathbf{x}^*) + \frac{\eta}{2}G^2. \tag{4}$$

Summing this up over all times gives

$$\sum_{t=0}^{T-1} f_t(\mathbf{x}_t) + (\Phi_T - \Phi_0) \le \sum_{t=0}^{T-1} f_t(\mathbf{x}^*) + \frac{\eta}{2}G^2 T.$$

Now using that $\Phi_T \ge 0$ and $\Phi_0 = D^2/(2\eta)$ completes the proof.

To prove (4), let's calculate

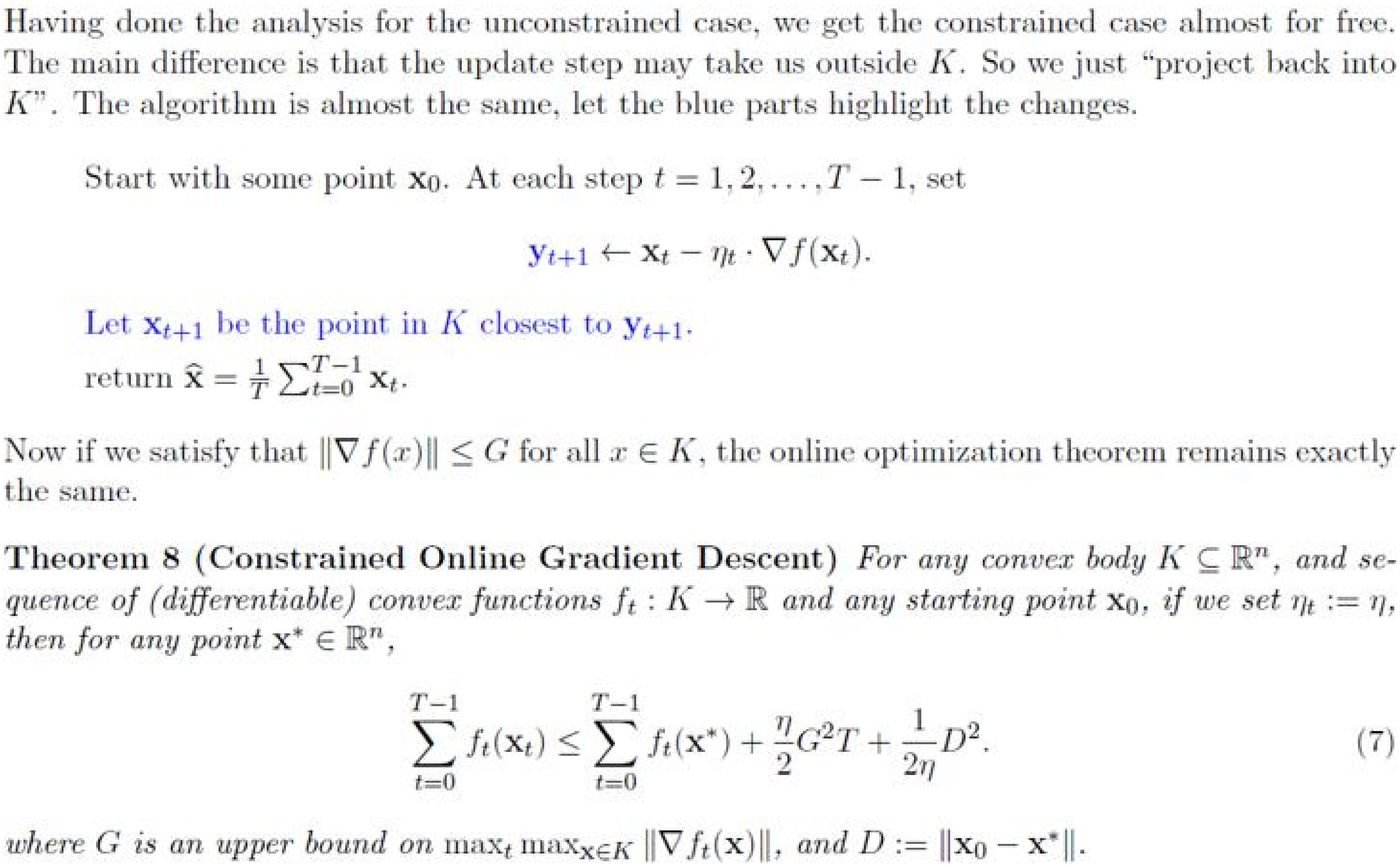
$$\begin{aligned}
\Phi_{t+1} - \Phi_t = \frac{1}{2\eta}\left(\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 - \|\mathbf{x}_t - \mathbf{x}^*\|^2\right) &\overset{(F1)}{=} \frac{1}{2\eta}\left(\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + 2\langle \mathbf{x}_{t+1} - \mathbf{x}_t, \mathbf{x}_t - \mathbf{x}^*\rangle\right) \\
&= \frac{1}{2\eta}\left(\eta^2\|\nabla f_t(\mathbf{x}_t)\|^2 - 2\eta\langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^*\rangle\right) \\
&\le \frac{\eta}{2}G^2 - \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^*\rangle. \qquad (5)
\end{aligned}$$

Next we use the convexity of f (via Fact 3) to bound the difference

$$f_t(\mathbf{x}_t) - f_t(\mathbf{x}^*) \le \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^*\rangle \tag{6}$$

Summing up (5) and (6) means the inner-product term cancels, and gives us the amortized cost bound (4), and hence proves Theorem 7. ■

Constrained Minimization



Having done the analysis for the unconstrained case, we get the constrained case almost for free. The main difference is that the update step may take us outside K. So we just "project back into K". The algorithm is almost the same, let the blue parts highlight the changes.

Start with some point $\mathbf{x}_0$. At each step $t = 1, 2, \ldots, T-1$, set

$$\mathbf{y}_{t+1} \leftarrow \mathbf{x}_t - \eta_t \cdot \nabla f(\mathbf{x}_t).$$

Let $\mathbf{x}_{t+1}$ be the point in K closest to $\mathbf{y}_{t+1}$.

return $\widehat{\mathbf{x}} = \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{x}_t$.

Now if we satisfy that $\|\nabla f(x)\| \leq G$ for all $x \in K$, the online optimization theorem remains exactly the same.

Theorem 8 (Constrained Online Gradient Descent) *For any convex body $K \subseteq \mathbb{R}^n$, and sequence of (differentiable) convex functions $f_t : K \to \mathbb{R}$ and any starting point $\mathbf{x}_0$, if we set $\eta_t := \eta$, then for any point $\mathbf{x}^* \in \mathbb{R}^n$,*

$$\sum_{t=0}^{T-1} f_t(\mathbf{x}_t) \leq \sum_{t=0}^{T-1} f_t(\mathbf{x}^*) + \frac{\eta}{2} G^2 T + \frac{1}{2\eta} D^2. \tag{7}$$

where G is an upper bound on $\max_t \max_{\mathbf{x} \in K} \|\nabla f_t(\mathbf{x})\|$, *and* $D := \|\mathbf{x}_0 - \mathbf{x}^*\|$.

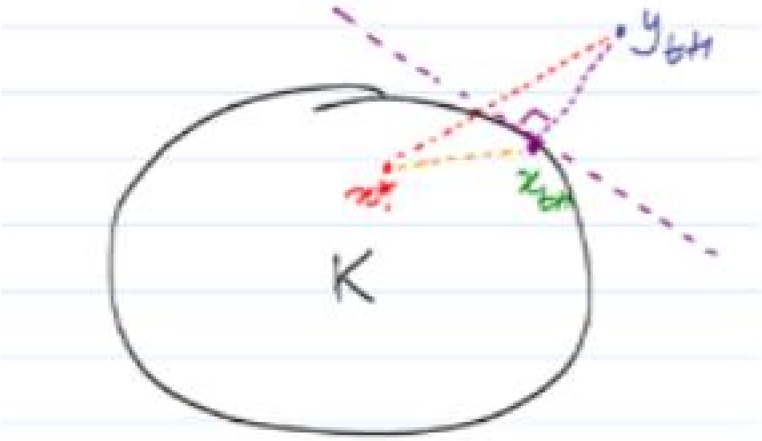


Figure 3: The projection ensures that $\mathbf{x}^*$ lies on the other side of the tangent hyperplane at $\mathbf{x}_{t+1}$, so the angle is obtuse. This means the squared length of the "hypotenuse" is larger than the squared length of either of the sides.

But now we claim that

$$\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 \leq \|\mathbf{y}_{t+1} - \mathbf{x}^*\|^2.$$

Indeed, since $\mathbf{x}_{t+1}$ is the "projection" of $\mathbf{y}_{t+1}$ onto the convex set K. The proof is essentially by picture (see Figure 3):

This means the changes die out almost immediately. Indeed,

$$\begin{aligned}\Phi_{t+1} - \Phi_t \leq \frac{1}{2\eta}\left(\|\mathbf{y}_{t+1} - \mathbf{x}^*\|^2 - \|\mathbf{x}_t - \mathbf{x}^*\|^2\right) &= \frac{1}{2\eta}\left(\|\mathbf{y}_{t+1} - \mathbf{x}_t\|^2 + 2\langle \mathbf{y}_{t+1} - \mathbf{x}_t, \mathbf{x}_t - \mathbf{x}^*\rangle\right) \\ &= \frac{1}{2\eta}\left(\eta^2\|\nabla f_t(\mathbf{x}_t)\|^2 - 2\eta\langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^*\rangle\right) \\ &\leq \frac{\eta}{2}G^2 - \langle \nabla f_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^*\rangle. \qquad (9)\end{aligned}$$